LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-421545

# A Method to Detect Differential Gene Expression in Cross-Species Hybridization Experiments at gene and probe level

Anu Chakicherla, James Felton, David Rocke, Ying Chen, REbecca Wu

December 17, 2009

Biomedical Informatics Insights

## Disclaimer

# A Method to Detect Differential Gene Expression in Cross-Species Hybridization Experiments at gene and probe level

Ying Chen[1£], Rebecca Wu[4], James Felton[4], David M. Rocke[2] and Anu Chakicherla[3£*]

[1] Department of Statistics, University of California Davis, Davis, CA

[2] Division of Biostatistics, University of California Davis, Davis, CA

[3] Computation Directorate, Lawrence Livermore National Laboratory, Livermore, CA

[4] Chemistry, Materials and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA

[*]To whom correspondence should be addressed

[£]Y.C. and A.C made equal contributions to this study

## ABSTRACT

**Motivation:** Whole genome microarrays are increasingly becoming the method of choice to study responses in model organisms to disease, stressors or other stimuli. However, whole genome sequences are available for only some model organisms, and there are still many species whose genome sequences are not yet available. Cross-species studies, where arrays developed for one species are used to study gene expression in a closely related species, have been used to address this gap, with some promising results. Current analytical methods have included filtration of some probes or genes which showed low hybridization activities. But consensus filtration schemes are still missing.

**Results:** We proposed a novel masking procedure based on currently available target species sequences to filter out probes and studied a cross-species data set using this masking procedure and gene-set analysis. Gene-set analysis evaluates the association of some priori defined gene groups with a phenotype of interest. Two methods, Gene Set Enrichment Analysis (GSEA) and Test of Test Statistics (ToTS) were investigated. The results showed that masking procedure together with ToTS method worked well in our data set. We also present results of an alternative way to study cross-species hybridization experiments without masking. We hypothesized that the multi-probes structure of Affymetrix microarrays makes it possible to aggregate the effects of both well-hybridized and poorly-hybridized probes to study a group of genes. The principles of gene-set analysis were applied to the probe-level data instead of gene-level data. The results showed that ToTS can give valuable information and thus can be used as a powerful technique for analyzing cross-species hybridization experiments.

**Availability:** Software in the form of R code is available at http://anson.ucdavis.edu/~ychen/cross-species.html.

**Contact:** chakicherla1@llnl.gov, lynchen@ucdavis.edu

**Supplementary Data:** Supplementary data are available at http://anson.ucdavis.edu/~ychen/cross-species.html

# 1. INTRODUCTION

Microarrays have become standard tools in biomedical and genomic research nowadays. Modulation of the expressions of thousands of genes simultaneously provides important insights into the molecular mechanisms of biological processes. However, in spite of the exponential growth in available whole genome sequences, there are still many organisms of interest, whose genomes have not been sequenced and therefore whose gene sequences are not all known. To study the gene expressions of these organisms, there are mainly two ways. The first way is to use cross-species hybridization, which is to hybridize the RNA samples of one species to the microarrays designed for a closely related species. The second way is to make customer-designed microarrays based on the currently available sequences of the organism being studied. Presently many biotech companies such as Affymetrix can provide such customer-designed services to fulfill the needs of genomic study. But this way may not be practical when the budget for the project is tight or the time is short or the sequences available are insufficient to be representative of the whole organism. In such cases, researchers often turn to the first way to solve the practical problem. Cross-species hybridization experiments usually cost less than making customer-designed arrays.

The cross-species approach has been employed in several studies in nonhuman primates, using human microarrays to analyze closely related species, such as chimpanzees, rhesus macaques and orangutans (Bigger *et al*., 2001; Huff *et al*., 2004), as well as more distantly related species, such as cattle, dogs, pigs or canines (Ji *et al*., 2003; Grigoryev *et al*., 2005). These studies assume that nucleotide sequence conservation within mammals is high enough to generate detectable signals. Despite the potential usefulness of cross-species hybridization studies, the quality of the gene expression measures obtained in this way is in question. Two important aspects of measurement quality of cross-species hybridizations have been examined: accuracy and reproducibility. Several studies have reported that the cross-species results are reproducible (Nieto-Diaz *et al*., 2007; Bar-Or *et al*., 2006). However, reproducibility does not assure that cross-species results can provide valid biological information. The accuracy aspect is more important.

Since the RNA samples of one species are hybridized to the arrays designed for another species, the sequence dissimilarity between the two species will cause the hybridization signals to be low compared to same-species hybridization (Bar-Or *et al*., 2006; Ji *et al*., 2003). The question has been raised whether cross-species hybridization studies are able to generate valid biological results similar to those obtained by same-species hybridization studies. It has been shown that cross-species hybridization can be used to detect within-species expression differences without discernible loss of information, as long as the two species are not too highly diverged (Oshlack *et al*., 2007). This provides an evidence of the validity of cross-species hybridization studies. However, it is generally agreed that the array sensitivity in cross-species studies decreases compared to that of same-species studies. This means cross-species analysis gives more false negatives, and thus the accuracy of the analysis decreases. To improve the array sensitivity, some researchers suggested a filtration approach called masking procedure (Ji *et al*., 2003; Grigoryev *et al*., 2005). That is, to install a mask to screen off poorly hybridized probes in Affymetrix arrays. The approach did improve the array sensitivity to some extents. But the problem is there are no

consensus filtration schemes available. Thus, it is needed to provide guidelines for the selection of masks or investigate if there is a better alternative.

Previous cross-species studies have used different microarray platforms, such as short oligonucleotide arrays, long oligonucleotide arrays, and two-color cDNA arrays. Affymetrix high-density oligonucleotide GeneChips (Affymetrix, Santa Clara, CA) is a kind of short oligonucleotide microarrays. In Affymetrix system, an mRNA molecule transcribed from a gene is represented by a probe set composed of 11-20 probe pairs. Each probe pair consists of a 25 bases long perfect match probe (PM) and a 25 base long mismatch probe (MM). The multi-probes structure of Affymetrix microarrays may have an advantage for cross-species analysis compared to other platforms such as cDNA microarrays (Ji *et al*., 2003). The reason is that the presence of multiple probes of each probe set may increase its probability to match with the target sequence, and thus make it possible to produce a good measure of its expression. In this paper, we will focus on the investigation of statistical methods to improve the sensitivity of cross-species analysis using Affymetrix GeneChips and the results of a cross-species data set will be discussed.

# 2. METHODS

## 2.1. Cross-species hybridization experiment

Data used in this study was generated at Lawrence Livermore National Lab from a cross-species hybridization experiment to study the heterocyclic amine food mutagen, 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP). PhIP is the most abundant of the carcinogenic heterocyclic amines found in well-cooked or over-cooked meat and fish. This compound has been shown to be a potent initiator and promoter of prostate and other cancers (Wu *et al*., 2000). In this study, a DNA repair-proficient Chinese hamster ovary (CHO) cell line (5P3R2) was exposed to PhIP, at a dose level of 0.4 micro ml, and RNA was harvested at different time points (2 hours, 4 hours, and 8 hours) after exposure. The PhIP-treated RNA samples and untreated reference samples were hybridized to Affymetrix's mouse GeneChip, MG U74Av2, respectively. Two technical replicates were available at each time point. Thus we had 12 arrays with 197993 probes that correspond to 12488 probe sets on each array. The aim of this study is to identify genes which respond to PhIP in cell line 5P3R2.

Using the same approach as other similar studies (Enard *et al*., 2002; DE Moody *et al*., 2002; Ji *et al*., 2003), the current study also assumes the divergence between mammals to be in the order of less than 100 million years. This also suggests that the conservation of protein function might ensure that sequence identity is also sufficiently conserved. The cumulative effect of such conservation across related proteins and sequences of interest would leverage the effort in this study to use gene sets to analyze gene expression data rather than the use of single gene data points.

### *2.1.1. Preparation of cDNA, RNA, hybridization to Affymetrix chips*

Purified total RNA was analyzed to ensure quality and quantity using the NanoDrop spectrophotometer and gel microelectrophoresis using Agilent's Bioanalyzer. Total RNA (5ugms) from all samples was labeled using a modified version of the Ebervine method. RNA was reverse transcribed using an oligo-dT T7 promoter primer and double stranded cDNA generated using the RiboAmp RNA Amplification Kit (Cat # KIT0209) according to manufacturer's directions as follows: a primer containing T7 RNA polymerase promoter sequence-oligodT was annealed to the total RNA molecules by heating at 65◦C for 5 minutes. A master mix including dNTPs was prepared using the RiboAmp kit and added and first strand cDNA generated by incubation at 42◦C for 45 minutes. The RNA is then degraded using a proprietary nuclease enzyme at 37◦C for 20 minutes followed by inactivation of nuclease at 95◦C, 5 minutes and cooled to 4◦C. Thereafter, the second strand cDNA synthesis was initiated by annealing another proprietary primer (Primer B) supplied with the RiboAmp kit to the first strand cDNA fragments (95◦C for 2 minutes, chill to 4◦C). Post addition of the second strand cDNA synthesis mix, the reaction is incubated at 25◦C for 5 minutes, 37◦C for 10 minutes and then finally at 70◦C for 5 minutes. The reaction mixture was then chilled to 4◦C and the double stranded cDNA thus generated was purified using columns and proprietary buffers supplied along with the RiboAmp kit with vendor provided purification protocol. The Enzo BioArray HighYield Transcript Labeling Kit (Cat # 42655-20) was used to generate labeled antisense RNA from the double stranded cDNA using T7 RNA polymerase and biotinylated UTP for amplification mediated labeling by incubation for 5 hours at 37◦C in a shaking water bath. Post labeling, the antisense RNA (ranging from 500-1800 nts) was first purified using the RNEasy procedure (Qiagen), fragmented as described in the Affymetrix Gene Expression Analysis Technical Manual (Affymetrix, Santa Clara, CA) and evaluated for quality and quantity by microchannel electrophoresis on the Agilent 2100 Bioanalyzer.

### *2.1.2. GeneChip staining, scanning, image processing*

Affymetrix MG U74Av2 gene chips were hybridized using 15 μg of fragmented complementary DNA followed by washing and staining in an Affymetrix Fluidics Workstation as described in the Expression Analysis Technical Manual (Affymetrix, Santa Clara, CA). Hybridized chips were scanned and signals were detected using an argon-ion laser scanner (Agilent Technologies, Palo Alto, CA). Microarray reports were generated to assess the hybridization quality and individual CEL files were used for data preprocessing.

### *2.1.3. Data preprocessing step*

The resulting images were processed to give a raw intensity value for each probe. Then probe level data need to be converted to expression values. There have been quite a few methods which achieve this goal such as MAS5.0 (Affymetrix, 2001), MBEI (Li and Wong, 2001a,b), RMA (Irizarry *et al*., 2003) and GLA (Zhou and Rocke, 2005). All these methods contain mainly three steps: 1) background correction, which refers to the adjustment intended to remove background noise; 2) normalization, which is a technique to reduce nonbiological variation in different arrays; 3) summarization, which gives an expression measure for each gene or probe set.

Among these, RMA expression measure has become a widely accepted method for Affymetrix GeneChips. In Zhou and Rocke (2005), they compared the performances of these preprocessing methods through two real data sets and concluded that GLA and RMA outperform the other methods. For our cross-species data set, we applied both GLA and RMA methods to background-correct and normalize the whole data set before we do any further analysis.

## 2.2. Masking Procedure

In Ji *et al*. (2003) and Grigoryev *et al*. (2005), they applied a masking procedure, which was to filter out the poorly hybridized probes in the data preprocessing step, to improve the sensitivity of cross-species analysis. However, the masking procedure is very empirical and both studies lack details explaining the reasoning of mask selection. We decided to use the currently available CHO sequences to choose masks. We searched GenBank for all the available CHO sequences and then used BLAST to match those CHO sequences to the probe sequences of Affymetrix's mouse gene chip. It turned out that 907 mouse probes are 100% matched to currently available CHO sequences. We hypothesized that if a known CHO sequence is 100% matched to a mouse probe, there is less chance of cross-hybridization in this probe. In other words, we should keep such probes in the data preprocessing step. Following this hypothesis, we created a ratio $r_m = \frac{n_m}{n_T}$

Where $n_m$ represents the number of remaining probes that are 100% matched to available CHO sequences and $n_T$ represents the number of total remaining probes after applying the mask. Intuitively thinking, a strict mask will lead to small value of $n_m$ and also small value of $n_T$. The goal is to mask off as many unmatched probes as possible while keeping most of the matched probes. Thus, a larger value of $r_m$ results in a better mask based on our hypothesis. Three groups of masks were selected for use: PM only, PM-MM, and PM/MM. In each group, three masking thresholds together with five masking stringencies were tested. The three masking thresholds are 25th, 50th and 75th percentile of the data set. The five masking stringencies are 8.33%, 25%, 50%, 75%, and 100%. The masking stringency 8.33% means that the probe is masked off if it does not meet the masking threshold in at least 8.33% of the 12 arrays, that is 1 array. Similarly, the masking stringency 100% means that the probe is masked off if it does not meet the masking threshold in all of the 12 arrays. We tested 45 masks and Supplementary Table 1 shows the details and results of all these masks. If the masking threshold is the same, a smaller masking stringency will lead to larger $r_m$ value. In addition, if the masking stringency is the same, a larger masking threshold will lead to larger $r_m$ value. Based on these findings, we selected three best masks using the $r_m$ value in the three groups: PM only, PM-MM, and PM/MM and used them for further analysis.

## 2.3. Gene-set Analysis

In DNA microarray studies, single-gene analysis has some limitations. A successful microarray experiment can result in a long list of differentially expressed genes which may not be easy to be interpreted by biologists. On the other hand, no single gene may be detected if the change of expression is very moderate. Gene-set analysis can generally overcome these limitations to some extents. Quite a few statistical methods have been proposed in recent years to study gene sets (Barry *et al*., 2005; Subramanian *et al*., 2005; Kim *et al*., 2005; Rocke *et al*., 2005). The basic idea of gene-set analysis is to look at the expression patterns in a group of genes to find out if they are associated with a class label or differentially expressed under different experimental conditions. Usually the genes in a predefined gene set have some biological themes, such as coming from the same biological pathway or having similar cellular functions. Thus the results of gene-set analysis are much easier to interpret and can help biologists understand some fundamental biological mechanisms.

In our cross-species data set, if we preprocessed the probe-level data using a standard method such as MAS5.0 or RMA and fit a linear model for each gene, we found that no single gene met the threshold for statistical significance after adjusting for multiple hypothesis testing. This is due to the low sensitivity of cross-species data analysis. In Affymetrix systems, since each gene has multiple probes on the chips, it is very unlikely to see that all the probes match well with the target even if this gene is truly differentially expressed. The standard summarization method, such as MAS5.0 or RMA, which gives an expression measure of each gene based on the intensity of all its probes, will often lead to low expression measures for truly differentially expressed genes (Ji *et al*., 2003; Grigoryev *et al*., 2005). This makes it harder to distinguish between genes that are truly differentially expressed and genes that have minimal or no change at all.

Thus, we decided to use gene-set analysis for our data set. We combined masking procedure with gene-set analysis at the first step. Since there is no consensus on the selection of a good mask, we proposed another way to apply gene-set analysis to cross-species data without using masking procedure, that is, to investigate the probe-level data instead of gene-level data. By applying the general ideas of gene-set analysis to probe-level data, we can aggregate the effects of multiple probes and statistically test if there is a general trend of expression changes in a group of genes under different experimental conditions.

Through Sections 2.4-2.5, we will introduce the basic algorithms of two gene-set analysis methods and use the following notations: S is a predefined gene set, N is the number of genes in a gene set S, L is a rank list of all the genes based on its association with class phenotypes, and r is the Pearson correlation.

## 2.4. Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a highly developed version of gene-set analysis (Subramanian *et al.*, 2005), which utilizes the Kolmogorov-Smirnov statistic to measure the degree of differential gene expression in a gene set across binary phenotypes. It ranks all the genes based on their association with a class phenotype and test whether the members of a predefined gene set are uniformly distributed throughout this list. The main steps of GSEA are:

1. Rank the genes or probe sets based on the correlation (or another metric) between their expression and the class phenotypes;
2. Compute an enrichment score ES by using a running-sum statistic. Start from the top of the rank list and let the running-sum be 0. Increase it if a gene in the gene set is encountered or decrease it otherwise. The enrichment score ES is calculated as the maximum deviation of the running-sum from zero;

$$ES(S) = \max_i \left| \sum_{g_j \in S, j \le i} \frac{|r_i|^p}{\sum_{g_j \in S} |r_j|^p} - \sum_{g_j \in S, j \le i} \frac{1}{N - |S|} \right| \quad (1)$$

3. Permute the class phenotypes and repeat steps 1-2. This generates a null distribution for the observed ES and the empirical, nominal p-value can be calculated.

If an entire database of gene sets is tested, a final step is added to adjust for multiple hypothesis testing. GSEA has been successfully applied to several cancer date sets in detecting significant biological pathways (Subramanian *et al.*, 2005).

## 2.5. Test of Test-Statistics (ToTS) method

Test of test-statistics (ToTS) method is another kind of gene-set analysis which also aims to detect significant biological pathways or any priori defined gene sets instead of individual genes. It has been successfully applied to an ionizing radiation study of prostate cancer where the patient variability in response to low doses of ionizing radiation (LDIR) creates substantial difficulties in detecting any differential gene expression (Rocke *et al*., 2005). The method consists of three steps:

1. For each gene or probe set in the group, conduct a statistical test of the hypothesis that it is differentially expressed;
2. Obtain the test statistics of all the genes or probe sets in the group and conduct a statistical test of hypothesis that there is a detectable up-regulating or down-regulating signal in the aggregate of them;
3. Assess the significance of the group of genes or probe sets by doing gene permutation.

In the ionizing radiation study, the effects of each gene in different patients are aggregated to test the significance of a gene set. In the cross-species hybridization experiment using Affymetrix GeneChips, we can aggregate the effects of multiple genes or multiple probes to test the significance of a gene set.

# 3. RESULTS

## 3.1. Gene Groups and Pathways

The website of GSEA provides a molecular signature database (MSigDB) that has a collection of gene groups and pathways from online pathway databases, publications in PubMed, knowledge of domain experts, and so on. Most of the gene group and pathway information was collected in human. We downloaded the annotation files for the human chip HG-U133A and the mouse chip MG U74Av2 from

the website of Affymetrix. In addition a linking table that has Orthologs/Homologs information and thus links probesets from mouse chip to human chip, was also downloaded from website of Affymetrix. Using this linking table which provides 1-to-1 relationship of human to mouse probesets, we created a collection of mouse gene groups and pathways. This resulted in 522 gene sets.

A subset of gene groups and pathways were further selected from the collection of 522 gene sets, on the basis of their relevance to cancer, signaling, oxidative stress which is believed to be one of the outcomes of exposure to PhIP, mitochondrial pathways, pathways involving p53, pathways pertaining to androgen and estrogen, and electron transport pathways. This resulted in 100 gene sets, which will be investigated in our analysis. The list of these 100 gene sets can be found in Supplementary Table 2.

### 3.2. Using Masking Procedure

Our fist step for statistical analysis was to get loess-normalized PM, PM-MM, and PM/MM intensities, respectively. As described in Section 2.2, we applied three different masking thresholds and five different masking stringencies with PM, PM-MM, and PM/MM data. The details are listed in Supplementary Table 1. We used the value of $r_m$ to select best masks, which resulted in, PM>5.83, PM-MM>34.58, and PM/MM>1.27. We denote these masks as mask1, mask2 and mask3. After we applied these three masks, single-gene analysis still did not give us any significant result. Thus we tried gene-set analysis. GSEA failed to identify any significant gene set at the threshold fdr<0.05 while ToTS identified some significant ones at the threshold fdr<0.05.

Out of the 100 gene sets that we selected, 9 of them were statistically significant between treatment and control after applying all the three masks, as shown in Table 1. The last 3 columns of Table 1 listed the fdr-adjusted p-values for using ToTS and 3 masks, respectively. Supplementary Figure 1 shows the overlap of significant gene sets identified by applying mask1, mask2, and mask3. The complete list of significant gene sets by applying the three masks are shown in Supplementary Table 3. We have found that applying masking procedure together with ToTS method could help us identify some significant pathways. We then posed the question whether we could achieve meaningful results without applying masking procedure. The next several sections show the results of applying gene-set analysis in probe-level data.

### 3.3. Application of GSEA to probe-level data

Our first step for statistical analysis is to background-correct and normalize the whole data set using GLA algorithm (Zhou and Rocke, 2005). Through Sections 3.4, we used the same GLA algorithm. In Section 3.5, we will discuss the results of using another preprocessing method RMA.

The GSEA method was developed to test the association of a pre-defined gene set or pathway with binary phenotypes. The first step of GSEA is to rank all the genes based on their association with the class labels. There are several metrics that can measure the association, such as Pearson's correlation, signal-to-noise ratio, two-sample t-test statistics and so on. Thus GSEA can be generalized to data other than binary phenotypes, as long as we can find a metric that measures the association of each gene with class labels in the data set. For each probe in our data set, we applied a two-way ANOVA model with treatment and time as the two factors, and tested if the expression of the probe is statistically different

between treatment and control samples. Thus we obtained a t-score for each probe out of 197993 probes. We used the absolute values of these t-scores to rank all the probes instead of Pearson correlation or other statistics. For each pathway, we calculated an enrichment score based on Equation (1). Instead of using sample permutation, we used gene permutation here to calculate the empirical p-values of each pathway. That is, to randomly select genes as the sampling units and generate a distribution for the ES test statistic. It is criticized that gene permutation tends to give anti-conservative results because the independence assumption between genes are usually unrealistic (Goeman and Buhlmann, 2007). However, in the cases when there are only a few arrays available, it is impossible to do a large number of sample permutations. Gene permutation is often the alternative in such cases. In this cross-species data set, we only had 4 arrays available at each time point. Thus we used gene permutation to test the significance of pathways. In order not to worsen the violation of gene-gene independence in a gene set, we randomly selected genes instead of probes and keep each gene's probe structure intact.

We found that none of the pathways met the statistical significance level at 0.05. Figure 1 shows the histogram of fdr adjusted p-values for all the 100 pathways tested. Most of the pathways have p-values greater than 0.50.

### 3.4. Application of ToTS to probe-level data

Similar to GSEA, our first step using ToTS is to get a t-score for each probe by fitting a two-way ANOVA model. To test whether a pathway is associated with the treatment PhIP, we aggregated the t-scores of all the probes corresponding to the genes in this pathway, and performed a one-sample t-test and a one-sample Wilcoxon test. We hypothesize that there are effects in at least some genes of a pathway, but it may be reflected in only a small number of probes that correspond to the genes because of the poor hybridization qualities of cross-species data set. We expect to see that the t-scores would be biased in a positive or negative direction if there is such kind of diffuse response. Figure 2 shows the histogram of the t-scores of all the probes corresponding to pathway "PENG GLUCOSE DN". There is a slight upward trend of all the t-scores. We would like to verify that this trend is not seen by chance. Using the one-sample t-test or one-sample Wilcoxon test, we can test whether the mean or median of the collection of these t-scores are different from 0. In order to reduce the bias introduced here, we also performed gene permutation to generate a null distribution for the t-test statistic or Wilcoxon-test statistic and thus obtained empirical p-values for each pathway tested. The standard p-values by t-test or Wilcoxon-test correspond well to the empirical p-values. So we omit these results here. In addition, Wilcoxon-test with gene permutation tends to be more sensitive than t-test in this case. Only the results of empirical p-values by Wilcoxon-test are showed in Table 2. Six pathways were found to be statistically significant at the threshold 0.05. For a comparison of GSEA and ToTS, we also listed the p-values of these six pathways by GSEA in Table 2. None of them met the statistical significance threshold.

### 3.5. Robustness of using different preprocessing algorithms

Since there are quite a few preprocessing algorithms available, we want to see if they have a great influence in the cross-species studies. In Sections 3.3-3.4, we have used GSEA and ToTS method to the probe-level data preprocessed by GLA algorithm, it seemed that ToTS is more sensitive and gives better results than GSEA. Since RMA has been a standard and quite accepted method for Affymetrix

GeneChips, we also used RMA to preprocess the raw data and applied similar gene-set analysis methods. In Zhou and Rocke (2005), they showed that GLA has comparable performance with RMA. As expected, we identified 6 six pathways with fdr adjusted p-values less than 0.05. They are exactly the same as the ones identified using GLA algorithm. If we compare Table 3 and Table 2, the conclusion based on using GLA and RMA methods are the same. This provides evidence that preprocessing algorithms do not have great effects on the results here.

## 4. DISCUSSION

In this study, we have proposed a novel masking procedure based on currently available target species sequences to filter out probes and combined this masking procedure with gene-set analysis. We have also proposed an alternative for analyzing cross-species studies using Affymetrix GeneChips, that is, to apply the ideas of gene-set analysis to the probe-level data. Two gene-set analysis methods and their application to a cross-species data set were investigated. The results showed that ToTS has the better performance than GSEA.

The results from ToTS with the masking procedure gave 42 pathways with significant p-values by at least one mask, and of these 9 had significant p-values with all three masks used (Supplemental Table 3). A particularly relevant biological aspect of this result is that there is a significant overlap between these results and pathways contributing to the Cancer pathways (subway map of cancer pathways). These include the WNT signaling pathways, GSK3, B-catenin, eIF4, TERT, PI3K, and p53 among others. Although some of these are lost upon correcting for multiple hypothesis testing, the results are interesting, nevertheless, and deserve further analysis.

The results from ToTS at the probe-level analysis done without masking gave 6 significant pathways with p-values less than 0.05. An important detail from this analysis is that 4 of the 6 pathways are also shared by the 9 pathways identified by applying masking procedure: PENG GLUCOSE DN, PENG GLUTAMINE DN, PROTEASOME DEGRADATION, and PROTEASOMEPATHWAY. This consistency may be evidence of improved sensitivity of ToTS methods in analysis of cross-species data either at probe-level or gene-level.

The application of GSEA either at probe-level or gene-level of our cross-species data set didn't give any significant result. A problematic issue of GSEA is that the calculation of enrichment score of a particular gene set not only depends on the expressions of genes in this gene set, but also those outside of the gene set. When a particular gene set is biologically meaningful, the expressions levels of other genes should not affect the inference on this gene set. This problem was also reported in Dinu *et al*. (2007) and it may cause the power of GSEA to be low in some situations. Different from GSEA, ToTS uses t-statistics or Wilcoxon-statistics which does not depend on the expressions of other genes outside of a particular gene set. ToTS may be expected to have more power than GSEA. A comprehensive comparison between the two methods deserves further study.

Through the study of this cross-species data set, we have seen that ToTS successfully improved the sensitivity of cross-species analysis while GSEA failed to do so. ToTS have better performance than GSEA

here and thus is potentially a useful tool for cross-species studies. However, a comprehensive guideline for its usage in cross-species hybridization experiments is still necessary in order for it to gain widespread success.

## FUNDING

This work is supported by grants from the Department of Energy (DE-FG02-07ER64341), Air Force Office of Scientific Research (FA9550-07-1-0146), National Human Genome Research Institute (R01-HG003352), National Institute of Environmental Health Sciences Superfund (P42-ES04699), and NIH-NCI (CA5586112S1). It was performed in part under the auspices of the U.S. DOE by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

## REFERENCES

Affymetrix (2001) Microarray Suite User Guide, Version 5, Affymetrix.

Al-Shahrour,F. et al. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, *Bioinformatics*, 21, 2988-2993.

Anderson,P.W. et al. (2006) Cross-species hybridization of woodchuck hepatitis virus-induced hepatocellular carcinoma using human oligonucleotide microarrays, *World J. Gastroenterology*, 12, 4646-4651. Bar-Or,C. et al. (2006) Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results, *BMC Genomics*, 7, 110. Bar-Or,C. et al. (2007) Cross-species microarray hybridizations: a developing tool for studying species diversity, *Trends in Genetics*, 23, 200-207.

Barry,W.T. et al. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach, *Bioinformatics*, 21, 19431949.

Bigger,C.B. et al. (2001) DNA microarray analysis of chimpanzee liver during acute resolving hepatitis C virus infection, *Journal of Virology*, Aug, 7059-7066. Bigger,C.B. et al. (2004) Intrahepatic gene expression during chronic hepatitis C virus infection in chimpanzees, *Journal of Virology*, Dec, 13779-13792. DE Moody,D., McIntyre,L. (2002) Cross-species hybridisation of pig RNA to human nylon microarrays, *BMC Genomics*, 3, 27.

Dinu,I. et al. (2007) Improving GSEA for analysis of biologic pathways for differential gene expression across a binary phenotype, *COBRA Preprint Series*, Article 16.

Enard,W. et al. (2002) Intra-and interspecific variation in primate gene expression patterns, *Science*, 296, 340-343.

Goeman,J.J., Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues, *Bioinformatics*, 23, 980-987.

Grigoryev et al. (2005) In vitro identification and in silico utilization of sequence similarities using GeneChip technology, *BMC Genomics*, 6, 62.

Huff,J.L. et al. (2004) Gastric transcription profile of helicobacter pylori infection in the rhesus macaque, *Infection and Immunity*, Sept, 5216-5226.

Irizarry,R.A. et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4, 249-264.

Ji,W. et al. (2003) A method for cross-species gene expression analysis with high-density oligonucleotide arrays, *Nucleic Acids Research*, 32, e93.

Kim,S.Y., Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment, *BMC Bioinformatics*, 6, 144.

Li,C., Wong,W.H. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc. Natl. Acad. Sci. USA*, 98, 31-36.

Li,C., Wong,W.H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues, and standard error application, *Genome Biol*, 2, RESEARCH0032.

Nieto-Diaz,M. et al. (2007) Cross-species analysis of gene expression in non-model mammals: reproducibility of hybridization on high density oligonucleotide microarrays, *BMC Genomics*, 8, 89.

Oshlack,A. et al. (2007) Using DNA microarrays to study gene expression in closely related species, *Bioinformatics*, 23, 1235-1242.

Rocke,D.M. et al. (2005) A method for detection of differential gene expression in the presence of inter-individual variability in response, *Bioinformatics*, 21, 3990-3992.

Subramanian,A., Tamayo,P. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, 102, 15545-15550.

Tusher,A.G., Tibshirani,R., Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *PNAS*, 98, 5116-512.

Vahey (2003) Patterns of gene expression in peripheral blood mononuclear cells of rhesus macaques infected with SIVmac251 and exhibiting differential rates of disease progression, *AIDS Research and Human Retroviruses*, 19, 369-387.

Walker (2006) Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates, *Journal of Neuroscience Methods*, 152, 179-189.

Wu,R.W. et al. (2000) Genetically modified Chinese hamster ovary cells for investigating sulfotransferase-mediated cytotoxicity and mutation by 2amino-1-methyl-6-phenylimidazo[4,5-b]pyridine, *Environmental and Molecular Mutagenesis*, 35, 57-65.

Zhou,L., Rocke,D.M. (2005) An expression index for Affymetrix GeneChips based on generalized logarithm, *Bioinformatics*, 21, 3983-3989.

**Table 1.** Significant pathways identified by applying 3 masks with ToTS method. The last three columns show the FDR-adjusted p-values by using 3 masks and ToTS, respectively.

| Biological Pathway | No. of genes | Mask1 | Mask2 | Mask3 |
|---|---|---|---|---|
| HUMAN MITODB 6 2002 | 428 | 0.0091 | 0.0357 | 0.0091 |
| MRNA PROCESSING | 47 | 0.0214 | 0.03 | 0.0263 |
| PENG GLUCOSE DN | 157 | 0.0091 | 0.0091 | 0.0091 |
| PENG GLUTAMINE DN | 313 | 0.0091 | 0.0091 | 0.0176 |
| PENG LEUCINE DN | 180 | 0.0091 | 0.0111 | 0.0176 |
| PENG RAPAMYCIN DN | 229 | 0.0091 | 0.0091 | 0.0091 |
| PROTEASOME DEGRADATION | 32 | 0.0091 | 0.0091 | 0.0091 |
| PROTEASOMEPATHWAY | 21 | 0.0091 | 0.0091 | 0.0091 |
| PYK2PATHWAY | 29 | 0.03 | 0.0467 | 0.0176 |

**Table 2.** FDR-adjusted p-values of six pathways by ToTS and GSEA (data preprocessed by GLA).

| Biological Pathway | No. of genes | FDR_adjusted p-values by ToTS | FDR_adjusted p-values by GSEA |
|---|---|---|---|
| NELSON ANDROGEN UP | 86 | 0.033 | 0.20 |
| PENG GLUCOSE DN | 157 | 0.033 | 0.40 |
| PENG GLUTAMINE DN | 313 | 0.033 | 0.52 |
| MAPK CASCADE | 33 | 0.033 | 0.10 |
| PROTEASOME DEGRADATION | 32 | 0.033 | 0.39 |
| PROTEASOMEPATHWAY | 21 | 0.033 | 0.15 |

**Table 3.** FDR-adjusted p-values of six pathways by ToTS and GSEA (data preprocessed by RMA).

| Biological Pathway | No. of genes | FDR_adjusted p-values by ToTS | FDR_adjusted p-values by GSEA |
|---|---|---|---|
| NELSON ANDROGEN UP | 86 | 0.033 | 0.22 |
| PENG GLUCOSE DN | 157 | 0.040 | 0.31 |
| PENG GLUTAMINE DN | 313 | 0.040 | 0.47 |
| MAPK CASCADE | 33 | 0.033 | 0.21 |
| PROTEASOME DEGRADATION | 32 | 0.040 | 0.33 |
| PROTEASOMEPATHWAY | 21 | 0.033 | 0.20 |

**Figure 1.** The histogram of the fdr-adjusted p-values for the 100 pathways by GSEA.

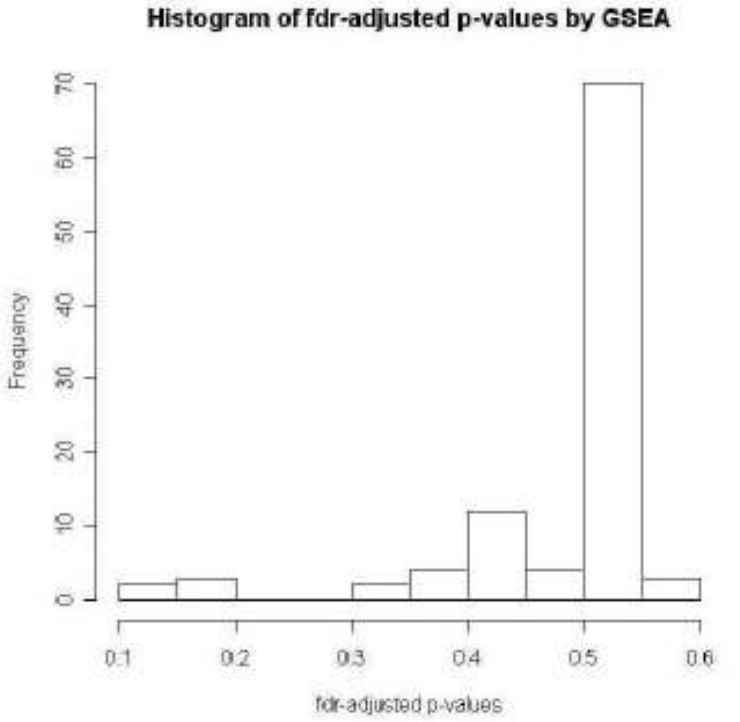



**Figure 2.** The histogram of the t-scores of all the probes corresponding to genes in gene set ”PENG GLUCOSE DN”. The right-sided vertical line is the median of all the t-scores.

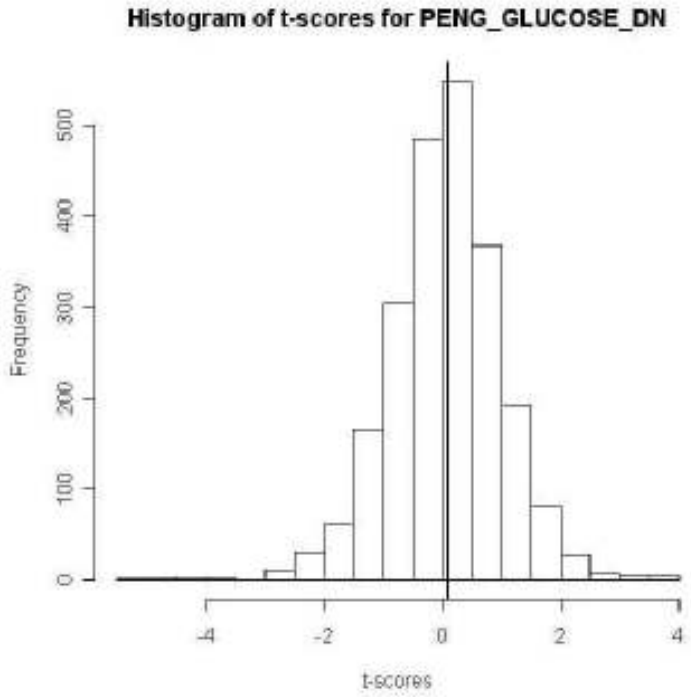
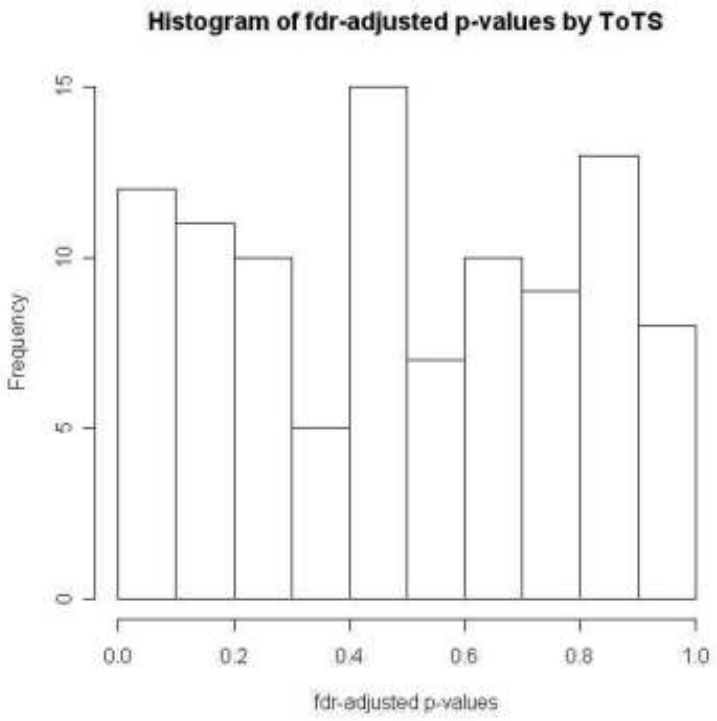


**Figure 3.** The histogram of the fdr-adjusted p-values for the 100 pathways by ToTS.




This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.